Biometrics by the Harbour, Hobart, 2015

How many letters are there in the alphabet?

Jeff Wood

The Australian National University

Uppercase Greek letters

Λ Ξ

A B

Bodybuilding competiton on Crete

MR KPHTH

Other use of non-Greek letters

DVD

BMW

Car number plates appear to only have letters from the intersection of the Greek and Latin alphabets.

For example KNK 6949

Is there a way of finding out how many letters are actually used by observing a sample of plates?

In general suppose that we have a population of objects which can be partitioned in to S classes, where S is unknown

How can we estimate S.

An example is the species richness problem - in a large population of animals how many distinct species are there?

Other examples - estimating the size of an author's vocabulary - estimating how many distinct people invest on the ASX, etc.

Alan Turing I.J. Good - Bletchley Park 007 IJG

Issues

Is the population finite or infinite?

Are objects sampled with or without replacement?

Are the objects arranged in groups (capture-recapture)

Do we want to use a nonparametric or a parametric estimate?
In other words are we happy to make assumptions about the relative abundances of different classes

Are we happy just to have a lower bound for the estimate of C? In practice an estimate of the lower bound.

Is it worthwhile to examine more objects (if this is expensive)

The number plate example is a good test bed.

We do not expect there to be a large number of rare letters, and we can find out the correct answer

Typical ACT car registration plate

YFV 54M

Observed 26 plates while travelling from Fisher to Coolamon Court (not very random). Maybe vanity plates are more common in some suburbs than others

0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G
8	4	5	2	3	4	5	6	6	9	5	6	7	1	7	4	3
H	I	J	L	M	N	P	Q	R	S	T	U	V	W	X	Y	Z
6	2	7	2	2	3	2	1	2	1	2	1	3	3	4	24	1

34 letters and digits observed (O and K missing)

The frequency of frequencies table is

0	1	2	3	4	5	6	7	8	9	24
?	5	7	5	4	3	4	3	1	1	1
f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{24}

Chao (1984) estimate of lower bound for f_0 is

$$\frac{f_1^2}{2f_2}$$

Lanumteang and Bohning (2012)

$$\frac{3 * f_1^3 f_3}{4 f_2^3}$$

Results

Method	Estimate	"Reliability"
chao1984	36	34-46
Lanumteang and Bohning (2012)	35	
ChaoLee1992	36	34-43
ChaoBunge	36	34-43
jackknife	39	33-45
unpmle	36	34-40
pnpmle	36	31-44
pcg	37	

Jackknife estimate of order 1 is

$$\widehat{N}_{J1} = d_n + (n-1)f_1/n$$

Jackknife estimator of order k is

$$\widehat{N}_{Jk} = d_n + \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} f_j$$

ChaoLee1992 and ChaoBunge are based on the idea of coverage

If the proportion of species i in the population is p_i the coverage, C, is $\sum_i p_i I(x_i > 0)$ where x_i is the number of individuals of species i observed.

$1 - C$ is the probability that the next object sampled is from a previously unobserved class

Other estimates assume that the number of individuals of each species observed is from a Poisson distribution. Assumptions are then made about the distribution of the parameters for the various species, e.g pcg stands for Poisson-compound gamma.

Using the first ten plates only, 28 digits and letters were observed. The estimates were

Method	Estimate	"Reliability"
chao1984	37	30-62
ChaoLee1992	37	30-56
ChaoBunge	36	30-61
jackknife	40	30-50
unpmle	36	
pnpmle	37	
pcg	41	


Заливот на коските
Bay of the Bones

Observed 52 plates in Macedonia

0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
17	18	18	14	13	16	15	16	23	20	14	5	5	3	4	4
G	H	I	J	K	L	M	N	O	P	R	S	T	U	V	Z
4	3	3	2	50	1	3	4	8	5	5	51	6	9	17	1

32 letters and digits observed (Q, W, X, Y missing)

The frequency of frequencies table is

0	1	2	3	4	5	6	8	9	13	14	15	16	17	18	20	23	50	51
?	2	1	4	4	4	1	1	1	1	2	1	2	2	2	1	1	1	1

Results

Method	Estimate	"Reliability"
chao1984	34	32-54
Lanumteang and Bohning (2012)	56	
ChaoLee1992	33	31-35
ChaoBunge	33	31-35
jackknife	39	33-45
unpmle	32	
pnpmle	33	
pcg	32	

Acknowledgments

Bob Anderssen and Bob Forrester